CO3: Low-resource Contrastive Co-training for Generative Conversational Query Rewrite

Yifei Yuan[1], Chen Shi[2], Runze Wang[2], Liyi Chen[3], Renjun Hu[2], Zengming Zhang[2], Feijun Jiang[2] and Wai Lam[4]
[1] University of Copenhagen, [2] Alibaba Group,
[3] Nankai Unversity, [4] The Chinese University of Hong Kong
yiya@di.ku.dk, liyichen@mail.nankai.edu.cn, wlam@se.cuhk.edu.hk
{deling.sc, yunze.wrz, renjun.hrj, zengming.zhangzm, feijun.jiangfj}@alibaba-inc.com

Abstract

Generative query rewrite generates reconstructed query rewrites using the conversation history while rely heavily on gold rewrite pairs that are expensive to obtain. Recently, few-shot learning is gaining increasing popularity for this task, whereas these methods are sensitive to the inherent noise due to limited data size. Besides, both attempts face performance degradation when there exists language style shift between training and testing cases. To this end, we study low-resource generative conversational query rewrite that is robust to both noise and language style shift. The core idea is to utilize massive unlabeled data to make further improvements via a contrastive co-training paradigm. Specifically, we co-train two dual models (namely Rewriter and Simplifier) such that each of them provides extra guidance through pseudo-labeling for enhancing the other in an iterative manner. We also leverage contrastive learning with data augmentation, which enables our model pay more attention on the truly valuable information than the noise. Extensive experiments demonstrate the superiority of our model under both few-shot and zero-shot scenarios. We also verify the better generalization ability of our model when encountering language style shift.

Keywords: Conversational Query Rewrite, Co-training, Low-resource Generation

1. Introduction

Recent progress in deep learning NLP techniques has witnessed a resurgent interest in developing conversational IR systems (Reddy et al., 2019; Choi et al., 2018a). Among these tasks, conversational query rewrite (CQR) aims to convert an in-context query to a more explicit form given its context history (Elgohary et al., 2019; Su et al., 2019; Pan et al., 2019b). The rewritten query is semantically equivalent to the original one but can be understood without referring to the context. The main research challenge in the CQR system is that conversational queries are often very concise. Information omission such as coreference and ellipsis can often be observed, where concepts in previous turns are easy to be referred back or omitted. Specifically, in the CQR task, for original QA pairs in a conversation, a manually rewritten query is provided. For example, as shown in Table 1, for the second query $\mathcal{Q}_2$, the term *"her"* in the original query is resolved as *"Beyoncé"* in the rewrite. In the third turn, the omitted information after the term *"What else"* is completed after rewriting.

Table 1: An example of a CQR system. $\mathcal{Q}$, $\mathcal{Q}^*$, and $\mathcal{A}$ denote the queries, the corresponding rewrites and the answers. Red color denotes the coreference rewrite part and blue denotes the ellipsis rewrite part.

Conversation Contexts	
$\mathcal{Q}_1$	What can you tell me about ***Beyoncé's*** voice ?
$\mathcal{A}_1$	Her tone and timbre as particularly distinctive...
$\mathcal{Q}_2$	What are some other facts about *her* voice ?
$\mathcal{A}_2$	The New York Times commented her voice is "velvety yet tart"...
$\mathcal{Q}_3$	What else ?
$\mathcal{A}_3$	Other critics praises she was "capable of punctuating any beat".
Query Rewrites	
$\mathcal{Q}_2^*$	What are some other facts about *Beyoncé's* voice ?
$\mathcal{Q}_3^*$	What else *can you tell me about Beyoncé's voice* ?

To address the research challenges in the CQR task, generative CQR has gained great research interest recently, which aims to generate high-quality rewrites and formulates it as a standard text generation problem (Elgohary et al., 2019; Su et al., 2019). However, it has several drawbacks. First of all, traditional generative models often rely on a large amount of gold rewrite data, whose annotation process is often very expensive. In addition, existing few-shot based models are often vulnerable to the inherent noise due to limited data size. Since the quality of well labeled data is vital to the rewrite performance, how to reduce the impact of noise is an important yet underexplored problem. Furthermore, since different annotator writing styles may result in shifted data distribution, a performance degradation occurs when testing cases come from a different data source dissimilar to the training set (Hao et al., 2021).

In this work, we study the generative CQR task

Work done when the author was an intern at Alibaba.

LREC-COLING 2024, pages 3394–3406
20-25 May, 2024.

under low-resource scenarios. Since pre-trained language models have shown great few-shot and zero-shot learning abilities in many NLP tasks, we develop our model based on pre-trained GPT-2 (Radford et al., 2019). In order to better leverage the large amount of unlabeled data, we propose a co-training paradigm based on iterative pseudo-labeling. Specifically, we aim to train two separate models namely *Simplifier* and *Rewriter* together, where the Simplifier takes the rewritten query as input and outputs the original query and the Rewriter works on the other way round. Both warmed-up by a small amount of labeled data, in each iteration, the two models first make predictions on the unlabeled data, then the pseudo data generated by the Simplifier is used to train the Rewriter and vice versa. Our model leverages the dual nature of the two models and performs iterative training with only unlabeled data, which largely alleviates the heavy cost of obtaining the gold labeled data. Furthermore, by sampling outputs from one model and inputs generated from the other, the paradigm reduces the gap of the distribution between target and output results, thus equipping the model with enhanced robustness when tackling the noise shift in heterogeneous training and testing data.

To reduce the impact of noise in the input queries, we further enhance the model by employing a contrastive learning based data augmentation strategy. Inspired by Gao et al. (2021), we augment the input text by passing it to the model twice to obtain two different embeddings with the same dropout rate. We divide the contrastive loss into internal and external parts. The former considers the two augmented embeddings as positive pairs, while the latter takes the average of the two embeddings and the target embedding as positive pairs. This strategy involves more changes to the original data and helps to learn the common semantic features between the similar inputs and distinguish the differences between dissimilar ones.

We conduct extensive experiments on two datasets. Our model outperforms state-of-the-art methods under both few-shot and zero-shot settings. Furthermore, we investigate the effect of weakly labeled data size on the performance by adjusting the confidence thresholds and enlarging the unlabeled dataset. The results show that the performance can still be improved when the unlabeled dataset is large enough. In addition, to show that our model has better generalization ability than existing methods, we further perform cross training and testing among two datasets.[1]

In conclusion, the main contributions are: (1) We propose a novel framework for generative CQR tasks in low-resource settings. Our framework combines a Simplifier and a Rewriter through iterative pseudo-labeling, leveraging the contrastive co-training paradigm. (2) We employ an effective contrastive learning based data augmentation strategy to distinguish the truly valuable information from the noise in the input. (3) Extensive experiments and analyses are performed to show the effectiveness and the superior generalization ability of CO3 when encountering language style shift.

2. Related Work

2.1. Conversational Query Rewrite

CQR aims to generate explicit rewrites for abbreviated in-context queries (Tredici et al., 2021). Following this line, many efforts treat this task as a module of the conversational system, including performing query expansion that selects important terms in the history context (Voskarides et al., 2020; Mele et al., 2020), encoding the user's question in a latent space (Yu et al., 2021), and contextualizing query embeddings within the conversation (Krasakis et al., 2022; Lin et al., 2021), etc.

Aiming at generating rewrites that are clear to humans reader, generative CQR treats the task as a standard text generation problem which can be solved via a Seq2Seq model (Elgohary et al., 2019; Pan et al., 2019b; Su et al., 2019). Further improvements are made to make the generated rewrite more accurate by developing a multi-task framework (Rastogi et al., 2019; Song et al., 2020; Zhang et al., 2020), incorporating semantic knowledge (Xu et al., 2020; Hao et al., 2021; Liu et al., 2020), or adding multimodal information (Yuan et al., 2022). However, these works often rely on large amount of human rewrite data (Vakulenko et al., 2021b), whose annotation phase is very expensive. We focus on the generative query rewrite under the low-resource scenario. Under this setting, the work by Yu et al. (2020) is the most relevant, which proposes two methods named rule based and self-training to transform ad hoc search sessions as pseudo target query rewrites.

2.2. Co-training Paradigm

As an extension of self-training, co-training is a semi-supervised learning technique where two or more models are trained by each other's predictions (Blum and Mitchell, 1998; Abney, 2002). In NLP areas, Wan (2009) first proposes a co-training approach to make use of unlabeled Chinese data. Wu et al. (2018) focus on the selection of samples and employ a reinforcement learning method to learn a data selection policy with a small labeled dataset. Chen et al. (2018) co-train the

[1] The code is available in https://github.com/yfyuan01/CO3.

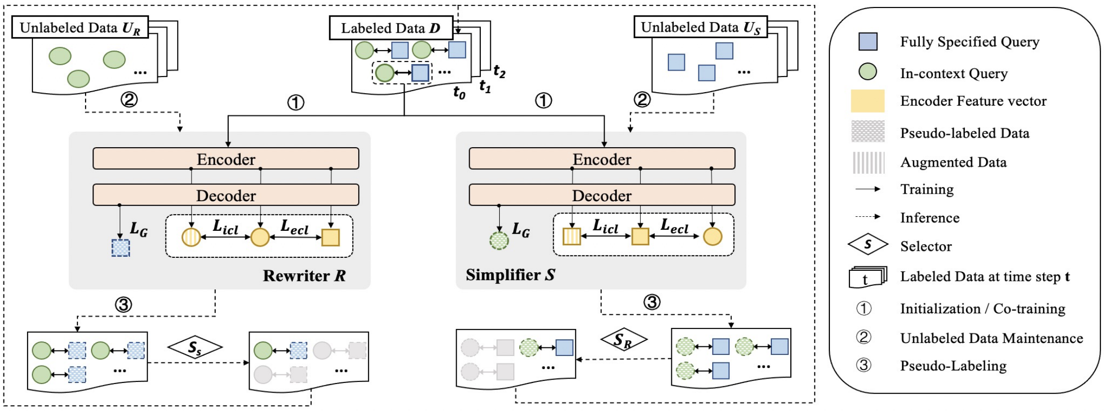

Figure 1: The overall framework of our proposed paradigm.

embeddings of knowledge graphs, whose performance improves at each iteration. In conversation-based tasks, co-training has been employed in the conversation disentanglement task (Liu et al., 2021b). Two neural modules called message-pair and session classifier are co-trained with pseudo data built from an unannotated corpus.

3. Our Framework

3.1. Problem Formulation

Conversational query rewrite aims to reformulate an in-context query to a more explicit form that can be understood without previous context history. Given a conversation context H with $m-1$ turns, it usually consists of several queries and can be denoted as $H = (q_1, q_2, ..., q_{m-1})$. Since the queries in the conversation often contain information omission, the task is to generate a rewrite q^* for the query on the latest turn q_m based on H. Specifically, in generative CQA, a query rewriter is trained to generate the de-contextualized rewrites given the conversation history of previous turns

$$q^* = Rewriter(H, q_m). \tag{1}$$

3.2. Framework Overview

Figure 1 depicts the overall structure of our framework. Our co-training framework consists of a *Rewriter* and a *Simplifier* with dual nature. The Rewriter generates the fully specified rewrite based on the original in-context query while the Simplifier works the other way round.

As shown in Figure 1, the whole paradigm is contained in a co-training loop where the Simplifier and Rewriter are trained together. The paradigm can be divided into three main steps. The first step ① is an initialization step where the Simplifier and Rewriter are warmed-up by a small number of labeled data D. At step ②, we maintain two unlabeled data pools, including the unlabeled Simplifier dataset U_S and the unlabeled Rewriter dataset U_R. After that, at step ③, the Simplifier and Rewriter predict and generate weakly labeled data on the unlabeled dataset respectively. The generated weakly labeled data is then fed into a Selector (S_S and S_R) which helps to filter the most confident subset of unlabeled samples for better training the models. In the next iteration, the filtered subset is then combined together with existing labeled data to form a synthetic dataset P and is augmented by a contrastive learning based strategy, where all the augmented data is later used to co-train the two models iteratively in the co-training step ①.

In order to enable our model to pay more attention on the truly valuable information, we enhance the model with a contrastive learning based data augmentation strategy, as shown in the yellow part within the Simplifier/Rewriter. Specifically, an in-batch contrastive loss is adopted which contains two parts. The internal part takes the two embeddings of the same sentence by feeding it to the encoder twice as positive pairs. The external part aims to shorten the pair-wise distance between model outputs and ground-truth rewrites while distinguish the differences between unpaired ones.

3.3. Co-training Paradigm

Model Initialization. We give a detailed description about the Simplifier and Rewriter in our framework. Both models can be initialized by generative models such as GPT-2 (Radford et al., 2019).

Simplifier is designed to transform the fully specified queries q^* into the simplified original in-context version q. Specifically, some terms or specific parts in the input queries may be replaced with pronouns or omitted in the rewrites. For example, after being simplified, the query "*What empires survived the Bronze Age collapse?*" is converted into "*What empires survived?*".

Rewriter learns to "put context back" to the contextual queries that contain coreference or ellipsis. It is the model we eventually wish to have and is the only model used during the inference stage.

In the few-shot setting, we use a small amount of well-labeled data to warm up and initialize the two models. Both Simplifier and Rewriter are initialized by the same set of data. For a small labeled dataset D, each data sample can be represented as (H, q, q^*), where H is the conversation history, q is the original query, q^* is the gold rewrite. The dataset can be directly used to train the Rewriter. By reversing the source and target query, the Simplifier can be trained inversely from q^* to q. In the zero-shot setting, since the gold labeled data is not available, we instead use some weakly labeled data to warm up and initialize the two models. The weakly labeled data is obtained by manually applying some pre-defined rules on the fully specified unlabeled dataset U_S. The rules are defined in the same way as (Yu et al., 2020), including replacing some noun phrases with pronouns, etc.

Unlabeled Data Maintenance. We maintain two additional unlabeled data pools including the unlabeled Simplifier dataset U_S and the unlabeled Rewriter dataset U_R. Each data sample in the unlabeled Simplifier dataset U_S is a user search log that contains several fully specified queries. In comparison, the unlabeled Rewriter dataset U_R contains real conversations where each query is contextual. Details of the datasets are described in Section 4.

Pseudo-labeling. After warming up the two models, the Simplifier and Rewriter then make predictions on unlabeled dataset U_S and U_R respectively. For each data input $q_s \in U_S$ and $q_r \in U_R$, the two models generate weakly labeled data as q'_r and q'_s. We then use a Selector to filter out predictions with low confidence by setting two confidence thresholds. We set the confidence score of the generated data as the generation likelihood score of both models. The two pseudo-labeled datasets are then fused together to form a synthetic dataset P for further training the model.

Model Co-training. The synthetic dataset P together with the labeled dataset D are used to train a better Simplifier and Rewriter model in the next iteration. Since the well-labeled data is limited and hard to obtain, we hope that the large amount of weakly labeled data helps the model learn the common patterns of the input queries. Before training the model, all the training data is augmented via a contrastive learning strategy which we will describe in detail in the next section. To avoid over-fitting, the two models are reinitialized in every iteration. For the Simplifier, we feed the well specified queries to obtain the abbreviated version and the Rewriter aims to put context back to provide the rewrites. After training, at the end of each iteration, the models of the next iteration will be overdriven by the newly trained models. Detailed algorithm is shown in Appendix A.

Generation Loss. The Simplifier and Rewriter both adopt the standard generation loss as the basic training loss. At each time step j, the decoder output is determined based on the generated sentence at the previous time steps $y_{<j}$. We minimize the negative log-likelihood of generating the target sentence y given the input x and context history H

$$L_G = \min \ -\sum_{j=1}^{|y|} logP_\theta(y_j|y_{<j}, x, H), \qquad (2)$$

where $|y|$ is the length of the generated sentence. For Simplifier, the generated y is the simplified original query q, while for Rewriter, y is the fully specified query q^*.

3.4. Contrastive Data Augmentation

We propose a simple but effective contrastive learning based data augmentation method. Motivated by SimCSE (Gao et al., 2021), we also pass the same input to the encoder twice to get two embeddings as positive pairs. Originated from the same sentence, the two embeddings differ in random dropout mask which can be seen as a special data augmentation form. After the dropout augmentation, a contrastive loss is added which takes two embeddings and the ground-truth rewrite embedding as input. Since the dropout can be viewed as a form of noise, this data augmentation technique helps the model learn the shared semantic pattern between the input sentences.

3.4.1. Internal and External Contrastive Loss

We divide the overall contrastive loss into internal and external parts. Both of them adopt the same in-batch contrastive loss function (details given in Section 3.4.2) that takes unpaired samples in a minibatch as negative pairs.

Internal contrastive loss. The internal contrastive loss aims to learn the common semantic features between the similar inputs. It takes the two augmented embeddings originated from the same sentence as positive pairs and aims to equip the model with better capacity to deal with noise. The process can be denoted as

$$L_{icl} = L_{cl}(Combine[\mathbf{Q}'; \mathbf{Q}'']), \qquad (3)$$

where $\mathbf{Q}'$ and $\mathbf{Q}''$ are two query embedding matrices from the same input by feeding into the encoder twice. The $Combine$ function is the concatenation of the two $N \times m$ embedding matrices into one $2N \times m$ matrix with an one-by-one manner. L_{cl} is the in-batch contrastive loss function.

External contrastive loss. The external contrastive loss focuses on shortening the distance between model outputs and the corresponding ground-truth rewrite. Therefore, it takes the average of the two sentence embeddings Q', Q'' and the target rewrite $\hat{Q}$ as positive pairs. Similarly, the external contrastive loss can be represented as

$$L_{ecl} = L_{cl}(Combine[AVG(\mathbf{Q}', \mathbf{Q}''); \hat{\mathbf{Q}}]), \quad (4)$$

We denote the final contrastive loss L_C as the sum of L_{icl} and L_{ecl}: $L_C = L_{icl} + L_{ecl}$.

3.4.2. Contrastive Loss Function

For the contrastive loss calculation, we follow the definition given in SimCLR (Chen et al., 2020), where the similarity between the representation of an input text and its corresponding positive pair is maximized and the similarity of in-batch unpaired instances is minimized. For a data sample in a minibatch with N instances and its augmented examples with the same size, the corresponding augmented data serves as the positive sample while the remaining $2N - 1$ data samples form the negative samples. The contrastive loss function in a minibatch can be represented as

$$l_{cl}(X_i, X_j) = \frac{exp(sim(X_i, X_j)/\tau)}{\sum_{k=1}^{2N} exp(sim(X_i, X_k)/\tau)}, \quad (5)$$

$$L_{cl}(X) = -\frac{1}{2N} \sum_{k=1}^{N} (l_{cl}(X_{2k-1}, X_{2k}) + l_{cl}(X_{2k}, X_{2k-1})), \quad (6)$$

where N is the batch size. X is an embedding matrix where the positive pairs in the batch are recorded one by one.

3.5. Training

The final loss combines the generation and contrastive loss

$$L_{all} = L_G + wL_C, \quad (7)$$

where w is the contrastive loss weight.

In order to distinguish the well-labeled and weakly labeled data, we also add a weight λ for the weak-labeled generation loss. The two types of data are combined by minimizing the loss function

$$L_G = L_G(D) + \lambda L_G(P), \quad (8)$$

$$L_{all} = L_G(D) + \lambda L_G(P) + wL_C. \quad (9)$$

4. Experiments

4.1. Datasets

In our work, both labeled and unlabeled data are used. The information of each dataset are presented in Table 2.

Table 2: The detailed information of the datasets used in our model. w/Omi. denotes if the dataset session contains information omission.

Name	CANARD	TREC	MS MARCO	QUAC
Session	304	50	9306	1000
Query	515	429	13799	7354
Labeled	Yes	Yes	No	No
w/Omi.	Yes	Yes	No	Yes

Labeled Dataset. We perform experiments on two different labeled datasets. First of all, we use the TREC CAst conversational search benchmark (Dalton et al., 2020). It contains 50 conversational sessions and 479 queries, each associated with a manual rewrite. In addition, we perform experiments on another query rewrite dataset named CANARD (Elgohary et al., 2019). To make it fit to the low-resource setting, we randomly sample 15% of the original dev set (originated from the QUAC training set) which contains 515 query-rewrite pairs.

Unlabeled Dataset for Simplifier. For Simplifier, the goal is to generate the simplified version of a fully specified query. We use the ad hoc search sessions collected from MS MARCO (Campos et al., 2016) directly. Based on the original MS MARCO QA dataset, the artificial search sessions are created using embedding similarity. Each query in the session is consistent with other queries in semantics without any information omission. We then filter the question-like search sessions from the original dev set and treat each session as a conversation. The total number of session is 9306 with 13799 different queries. One example session of the dataset is: *What is the australian flag? || What is the population of australia?*.

Unlabeled Dataset for Rewriter. For Rewriter, the queries must be context-aware and contain coreference and ellipsis. We use the Question Answering in Context (QUAC) dev dataset (Choi et al., 2018b) which perfectly fits to our setting. The final unlabeled dataset consists of 1000 unique sessions with 7354 queries in total. Each session and query have 440 and 6.5 tokens on average respectively. One sample session of this dataset is: *What is the australian flag? || What is the population of this country?*

4.2. Compared Methods

We compare our model with the following methods:

- **Original**. The rewrite is set to be the same as the input query.

Table 3: The experimental results on TREC dataset. For GPT-2 based models, we report both results of GPT-2 (base) (within the brackets) and GPT-2 (medium). * denotes that CO3 performs significantly better than other GPT-2 based baselines at 0.05 level using the two-tailed pairwise t-test. † denotes the upgraded L-CO3 outperforms all the baselines significantly.

	Model	**BLEU-1**	**BLEU-2**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**EM**	**NDCG@3**
	Original	72.50	66.17	79.71	65.66	79.66	18.65	30.40
	Allen Coref	79.37	74.29	86.04	76.72	85.94	36.13	43.59
Zero-shot	GQR	16.02	10.63	27.37	13.13	27.29	1.47	12.56
	GPT-2	15.41 (15.45)	10.54 (10.40)	27.17 (28.46)	12.42 (12.86)	26.75 (28.12)	1.17 (1.86)	11.32 (11.56)
	MS MARCO	35.19 (34.62)	19.90 (19.73)	31.06 (29.93)	13.18 (13.21)	30.41 (29.39)	0.93 (0.93)	16.90 (14.32)
	Rule Based	82.49 (79.31)	74.29 (72.30)	82.92 (82.93)	71.03 (70.53)	81.55 (81.86)	25.87 (26.81)	43.72 (43.25)
	CO3	**83.94*** (**80.91**)	**75.36*** (**73.37**)	**84.08*** (**83.08**)	**72.32*** (**71.31**)	**82.94*** (**82.02**)	**27.91*** (**27.04**)	**45.72*** (**44.67**)
	L-CO3	89.42†	77.31†	89.06†	74.90†	85.26†	30.55†	48.90†
Few-shot	Seq2Seq	72.11	62.47	78.75	65.61	78.02	6.45	20.42
	GQR	84.84	78.80	87.42	77.93	86.40	40.82	47.28
	GPT-2	84.61 (83.20)	78.62 (77.00)	87.27 (85.52)	77.86 (75.79)	86.25 (84.66)	40.79 (35.89)	46.74 (43.28)
	Rule Based	85.71 (82.35)	79.66 (76.23)	88.08 (85.91)	78.71 (75.97)	86.97 (85.09)	40.79 (36.13)	49.21 (46.76)
	Self-Learn	85.12 (**83.53**)	79.73 (77.51)	88.22 (86.82)	79.36 (76.90)	87.38 (85.91)	43.12 (38.23)	49.24 (46.53)
	CO3	**85.87*** (83.42)	**80.24*** (**78.14**)	**89.04*** (**86.95**)	**80.08*** (**77.48**)	**87.92*** (**86.36**)	**44.05*** (**40.79**)	**50.43*** (**48.26**)
	L-CO3	90.05†	86.47†	93.26†	85.28†	92.43†	49.07†	56.22†

Table 4: The experimental results of our model compared with the baseline models on CANARD dataset.

	Model	**BLEU-1**	**BLEU-2**	**BLEU-4**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**EM**
	Original	48.86	43.04	34.43	67.98	50.58	67.91	6.99
	Allen Coref	50.26	44.15	35.06	68.93	51.80	68.70	8.74
Zero-shot	GQR	9.07	5.64	2.34	15.83	4.61	14.79	0.18
	GPT-2	10.92 (11.99)	5.93 (6.81)	2.43 (3.08)	15.10 (16.70)	4.50 (5.48)	13.95 (15.46)	0.19 (0.39)
	MS MARCO	23.40 (23.42)	12.61 (11.95)	5.15 (4.44)	24.58 (23.29)	9.72 (8.87)	23.99 (22.70)	1.94 (1.03)
	Rule Based	52.25 (49.15)	41.83 (**40.71**)	29.53 (29.77)	56.43 (58.73)	39.40 (41.41)	54.57 (57.88)	3.11 (2.14)
	CO3	**53.21*** (**49.39**)	**42.73*** (40.31)	**30.80*** (**30.34**)	**59.25*** (**59.10**)	**42.27*** (**42.73**)	**58.40*** (**57.94**)	**3.74*** (**3.55**)
	L-CO3	58.63†	46.66†	37.57†	69.10†	52.01†	70.33†	9.47†
Few-shot	Seq2Seq	45.21	37.32	26.09	53.10	38.21	54.25	3.77
	GQR	48.03	41.20	30.98	56.72	44.10	58.82	7.84
	GPT-2	47.52 (47.23)	40.01 (38.80)	30.34 (28.50)	55.59 (54.34)	42.38 (38.92)	58.76 (54.34)	4.66 (3.88)
	Rule Based	55.05 (52.07)	46.72 (44.48)	35.70 (34.54)	65.36 (63.41)	48.66 (46.99)	64.40 (62.42)	7.96 (6.80)
	Self-Learn	55.77 (52.06)	47.40 (44.36)	36.15 (34.29)	65.84 (63.08)	48.86 (46.41)	64.75 (61.96)	7.57 (7.18)
	CO3	**57.55*** (**54.83**)	**48.55*** (**46.37**)	**36.94*** (**35.33**)	**66.59*** (**64.85**)	**49.35*** (**47.94**)	**65.68*** (**62.66**)	**9.02*** (**8.18**)
	L-CO3	64.29†	55.46†	41.73†	72.50†	55.28†	74.21†	12.33†

- **Allen Coref** (Gardner et al., 2018) is used for solving the coreference resolution problem in the query. We use it to generate query rewrites.

- **MS MARCO** fine tunes GPT-2 on the MS MARCO dataset via a language modeling task.

- **Seq2Seq** (Elgohary et al., 2019) is a neural Seq2Seq model where the encoder-decoder structure is based on bidirectional LSTM (Bahdanau et al., 2015; See et al., 2017).

- **GPT-2** (Radford et al., 2019) is adopted in both settings. In the few-shot setting, we fine-tune the model via cross-validation. In the zero-shot setting, we generate queries without any fine-tuning.

- **GQR** (Tredici et al., 2021) is a generative QR method based on T5-large (Raffel et al., 2020).

- **Rule-Based** (Yu et al., 2020) generates weakly labeled data by setting two simple rules, which create abbreviated query given its full version.

- **Self-Learn** (Yu et al., 2020) provides a method for generating the weakly labeled data. A Simplifier is trained separately and applied to the MS MARCO artificial sessions to generate weakly labeled data.

- **L-CO3** (Touvron et al., 2023) is an upgraded version of CO3, with the base model Llama, to test our model with the support of LLMs.

4.3. Experimental Settings

The code of our model is based on PyTorch and Huggingface Transformers (Wolf et al., 2019). In the few-shot setting, we fine-tune the model with 5-fold cross validation following (Yu et al., 2020). We split the sessions of two labeled datasets into five folds, where four are used for training and one is used for testing. With different training and testing portions, the whole process includes five rounds. We report the average score of them. Under the zero shot scenario, the whole dataset is used for the testing without splitting. The training is also conducted for 5 rounds with different random seeds. By default, we set the batch size as 4 and the learning rate as 5e-5. The evaluation metrics can be divided according to two aspects. We employ some

Table 5: Ablation study of our model on TREC dataset, where CL represents contrastive learning.

	Model	**BLEU-2**	**ROUGE-L**	**EM**
Few-shot	1.) w/o External CL	79.96	87.40	43.25
	2.) w/o Internal CL	80.03	87.44	43.69
	3.) w/o CL	79.72	87.34	43.19
	4.) w/o Simplifier	78.90	87.43	42.90
	CO3	80.24	87.92	44.05
Zero-shot	1.) w/o External CL	75.02	82.63	27.56
	2.) w/o Internal CL	74.88	82.31	26.80
	3.) w/o CL	74.30	81.61	25.90
	4.) w/o Simplifier	74.52	81.78	26.20
	CO3	75.36	82.94	27.91

generation evaluation metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and Exact Match (EM) to measure the rewrite quality. In addition, for the TREC CAST dataset, we also report the mean NDCG@3 to evaluate the ranking results with the rewritten query. In detail, the rewritten query is used to retrieve relevant passages with Anserini BM25 (Robertson and Zaragoza, 2009) toolkit and a BERT (Devlin et al., 2019) re-ranker is used to re-rank the candidates.

4.4. Main Experiment Results

Table 3 and Table 4[2] show the main experiment results, we have the following observations: first of all, pretrained language models have a great few-shot learning ability. Even with small amount of data, the Self-Learn model outperforms the Allen Coref model on the TREC dataset. In addition, in the few-shot setting, fine-tuning GPT-2 with small amount of labeled data has improved the BLEU-2 performance from 62.47 to 78.62 compared with traditional Seq2Seq model in TREC. Besides, under the zero-shot setting, directly using MS MARCO sessions to fine-tune GPT-2 model is far under satisfaction. However, by manually defining some rules, the Rule Based model achieves 82.49 and 52.25 BLEU-1 result in TREC and CANARD, which verifies the importance of the weakly labeled data. CO3 achieves the best overall performance among all the methods on two datasets in both settings. Specifically, the GPT-2 medium based version performs significantly better than other GPT-based baselines. In the few-shot setting, CO3 outperforms the Self-Learn method using the same amount of weakly labeled data. In the zero-shot setting, the superior results also prove the benefit of our co-training paradigm. Besides, with the help of LLM, the upgraded L-CO3 gains further performance lift. This demonstrates the superiority of our paradigm in the generative LLM era.

Table 6: Confidence threshold analysis of CO3.

	s_s	s_r	**BLEU-2**	**ROUGE-L**	**EM**
Few-shot	0	0	79.16	87.22	42.91
	50	70	77.69	86.68	41.78
	70	90	78.43	86.49	39.77
	90	**110**	**80.24**	**87.92**	**44.05**
	110	130	78.21	87.52	44.02
Zero-shot	0	0	74.07	81.70	20.28
	50	70	74.23	81.84	22.61
	70	**90**	**75.36**	**82.94**	**27.91**
	90	110	71.66	81.20	25.87
	110	130	74.16	80.83	23.08

Table 7: Dataset scale analysis of our model. Scale denotes the size of U_S and U_R datasets.

	Scale	BLEU-1	BLEU-2	ROUGE-L	EM
Few-shot	10k	85.64	79.26	87.81	44.11
	20k	85.91	79.98	87.99	44.20
	30k	86.17	80.34	88.05	44.35
	40k	**86.35**	**81.02**	**88.54**	**44.76**
Zero-shot	10k	83.83	75.47	81.98	26.11
	20k	84.20	76.23	81.71	26.47
	30k	84.90	76.79	82.97	27.98
	40k	**85.93**	**77.21**	**83.77**	**28.94**

4.5. Ablation Study

To evaluate the effect of different components of our framework, we report the performance of our model with several variants. According to Table 5, without the CL method, the performance decreases in both settings. Besides, contrastive loss is more effective in the zero-shot scenario than in the few-shot setting, which proves that using the contrastive learning based data augmentation technique helps to enhance the model when the gold data is not available. Among the two CL losses, the external loss plays a more important role, where the few-shot EM performance drops 0.8 percent without it. Furthermore, without the Simplifier, the few-shot and zero-shot EM performance decreases 1.15 and 1.71 percent, demonstrating the superiority of our co-training paradigm.

5. Extensive Analysis

To further evaluate the model capacity, we conduct several analysis to show the great potential and superior generalization ability of our model.

5.1. Weakly Labeled Data Scale Analysis

We first analyze the impact of the weakly labeled data involved in model training on the performance.

Confidence Threshold Analysis. We control the amount of training data by adjusting the two numbers and report the results under different Simplifier and Rewriter confidence thresholds. According to the results shown in Table 6, in TREC dataset, the

[2] Some numbers may be slightly different from the original paper because some evaluation codes are not publicly available. We use their model code and our own evaluation metric code to do the testing.

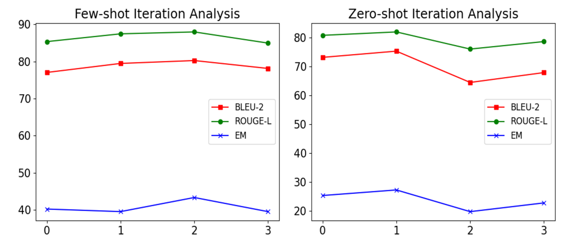

Figure 2: Performance of CO3 in each iteration.

best confidence threshold is lower in zero-shot than in few-shot. The result is because the quality of the generated data is worse in this setting and has a lower score. In addition, starting from zero when we enlarge the two thresholds, the performance first increases while starts to drop at certain stage. This is because when the confidence threshold is too small, large amount of noise data is introduced which negatively affects the rewrite quality. However, when the confidence threshold is too large, most weakly labeled data fails to participate in model training, thus causing model overfitting.

Dataset Scale Analysis. We fix the confidence thresholds and incrementally enlarge the unlabeled dataset by adding new instances. Table 7 shows the performance under different dataset scales. By enlarging the unlabeled dataset, the performance increases under both settings. Compared with the few-shot setting, dataset scale has more effect on the zero-shot scenario where gold data is unavailable. Notably, when the U_S and U_R dataset reach to 30k and 40k samples, the result exceeds the best performance (zero-shot EM increases from 27.91 to 28.94, few-shot EM increases from 44.05 to 44.76). This proves that we can further improve the performance by setting a high threshold with a larger dataset where large amount of high quality data is filtered to join model training.

5.2. Iteration Analysis

In Figure 2, under the few-shot setting, the model performance keeps increasing in the first three iteration and reaches to the best in the third iteration. In the zero-shot setting, the model takes less iterations to reach the peak. It demonstrates that without the gold labeled data, the model is more easy to suffer from overfitting. However, compared with traditional methods, our model prohibits data overfitting in two aspects. First, compared with methods with fixed dataset, new data can be introduced to the model in each iteration via additional weakly labeled data. Second, the dropout based data augmentation strategy makes sure that some random noise is added, ensuring that the data distribution is not strictly alike in each iteration.

Table 8: Generalization analysis on two datasets. T, C denote the TREC, CANARD dataset.

	Model	BLEU-1	BLEU-2	ROUGE-L	EM
T->C	Seq2Seq	35.92	24.12	43.97	2.03
	GPT-2	50.83	43.01	59.60	4.47
	Rule-Based	54.55	44.91	59.64	5.24
	Self-Learn	52.76	44.44	61.68	6.41
	CO3	**55.28**	**46.38**	**64.16**	**7.05**
C->T	Seq2Seq	68.87	59.34	75.23	5.37
	GPT-2	80.76	72.81	79.68	29.84
	Rule-Based	83.96	76.78	84.37	36.13
	Self-Learn	81.13	74.07	82.96	30.77
	CO3	**84.23**	**77.15**	**85.52**	**38.63**

5.3. Generalization Analysis

We train our models on one dataset while testing on the other to explore the generalization ability of our model. Table 8 shows that the traditional non-pretrained Seq2Seq model encounters severe performance drop when the testing data is different from the training data, where the TREC EM performance drops to 5.37 when training on the CANARD dataset, and the CANARD performance also decreases to 2.03 when training on TREC. This is mainly due to the writing style shift between the heterogeneous training and testing samples. Compared with raw GPT-2, models enhanced with weakly labeled data show better performance. This proves that large amount of weakly labeled data helps the model learn the common feature among queries that need to be rewritten. Our model achieves the best overall scores on all metrics concerning both two datasets, showing the superiority of CO3 on the cross-dataset robustness.

5.4. Loss Function Analysis

5.4.1. Weakly Labeled Data Weight Analysis

As shown in the upper part of Figure 4, when we increase the weakly labeled data weight λ, the performance first increases then decreases. The result verifies that although the quality of weakly labeled data may not be as good as the gold rewrite, it has a positive influence on the performance. Furthermore, it can be observed that under the zero-shot setting, the best weakly-labeled data weight is larger than what in the few-shot setting, where the model is more dependant on the large amount of unlabeled data for the lack of gold data for training guidance.

5.4.2. Contrastive Learning Weight Analysis

As shown in the lower part of Figure 4, we found that when the contrastive loss number w is around half of the generation loss (around 0.03 and 0.04), the model reaches the best score. We can also have the same observation as Section 5.4.1, that is, in the zero-shot setting, the contrastive loss (0.04) under the best performance is slightly larger

Index	Contexts	Current Query	Gold Rewrite	**Rule Based**	**Ours**
1	What are the different types of *sharks*? || Are *sharks* endangered? If so, which species? || Tell me more about *tiger sharks* || What is the largest ever to have lived on Earth? || What's the biggest ever caught? || What about for *great whites*? ||Tell me about *makos*.	What are their adaptations?	What are Mako shark adaptations?	Tell me about makos.	What are mako's adaptations?
2	What is *throat cancer*? || Is it treatable? || Tell me about *lung cancer*. || What are its symptoms? || Can it spread to the throat? || What causes *throat cancer*? || What is the first sign of it? || Is it the same as *esophageal cancer*?	What's the difference in their symptoms?	What's the difference in throat cancer and esophageal cancer's symptoms?	What's the first sign of esophageal cancer?	What's the difference in their symptoms?
3	What are the different types of *sharks*?	Are sharks endangered? If so, which species?	Are sharks endangered? If so, which species?	Are sharks endangered?	Are sharks endangered?
4	Tell me about *the Neverending Story film*. || What is it about? || How was it received? || Did it win any awards || Was it a book first?,	Who was the author and when what it published?	Who was the author and when was the Neverending Story film published?	Who was the author of the Neverending Story film and when what it published?	Who was the author of the Neverending Story film and when what the Neverending Story film published?
5	Tell me about the history of *toilets*. || Where does the term come from?	Where and when was the first invented?	Where and when was the first toilet invented?	Where and when was the first invented toilet?	Where and when was the first toilet invented?
6	Tell me about the history of *toilets*. || Where does the term come from? || Where and when was the first invented?	Why do the Brits call it a loo?	Why do the Brits call a toilet a loo?	Why do the Brits call a loo a loo?	Why do the Brits call a toilet a loo?
7	What is *the US Electoral College*? || How does it work? || Tell me about its creation || Why was the system chosen? || How has it changed over time? || What if *the electors* don't vote for *the pledged candidate*?	How has this changed election outcomes?	How have electors that don't vote for the pledged candidate changed election outcomes?	How has the US Electoral College changed election outcomes?	How has the electors changed election outcomes?

Figure 3: Real case and error case analysis. The first three are examples under the few-shot setting and the last four are under the zero-shot setting. The blue part denotes the resolved coreference or completed ellipsis in the gold rewrite. The red part denotes the errors in the model output.

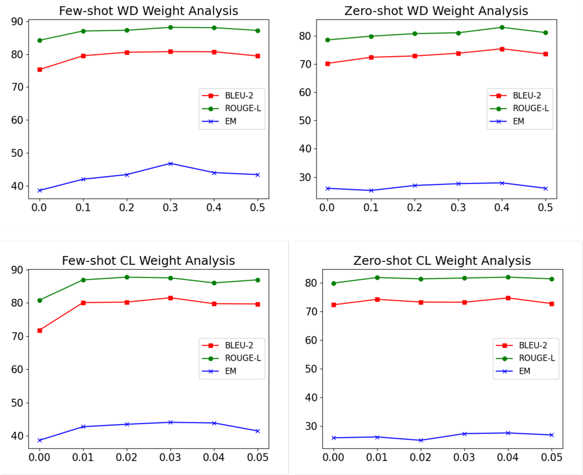

Figure 4: The upper part is the weakly-labeled data weight analysis and the lower part is the contrastive weight analysis.

than that in the few-shot setting (0.03). The result verifies that contrastive loss is more useful when there is few well-labeled data when training the model and can help the model better tackle the noise.

5.5. Real Case and Error Case Analysis

We provide some cases to compare CO3 with the rule-based model and analyze the potential drawbacks. As shown in Figure 3, CO3 has a better overall rewrite ability. For example, in case 1 and 2, the rule-based method outputs the rewrite of a wrong context sentence instead of the current query that needs to be rewritten. In case 6, the pronoun “it” in the original query should refer to “a toilet” in the context instead of “a loo”. However, some errors remain in both models. For instance, in case 3, although the rewrite seems easy and does not require any change, the last sentence is omitted by both models. Besides, as the conversation goes deeper, coreference that is challenging to both models is more common, such as coreference containing several entities (e.g. case 2) and coreference requiring reasoning between different entities (e.g. case 7).

6. Conclusion

We investigate the conversational query rewrite task under low-resource settings. We propose a co-training paradigm where a Simplifier and Rewriter are jointly trained. The Simplifier takes the fully specified query as input and outputs the abbreviated query and the Rewriter works the other way round. Based on iterative pseudo-labeling, the two models have dual nature where one takes the output from the other as input in each iteration. To distinguish the truly valuable information of the input, we enhance the model with a contrastive learning based data augmentation strategy. Experiments show the effectiveness of CO3 on two datasets. Extensive analyses are performed to prove the results can be further improved. Future works and limitations are discussed in Appendix B.

7. Ethics Statement

In adherence to ethical considerations, our work utilizes exclusively open-source datasets. We have strictly followed all licensing and intellectual property rights associated with these datasets.

8. Bibliographical References

Steven Abney. 2002. Bootstrapping. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 360–367.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Maria-Florina Balcan, Avrim Blum, and Ke Yang. 2005. Co-training and expansion: Towards bridging theory and practice. *Advances in neural information processing systems*, 17:89–96.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *IJCAI*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018a. Quac: Question answering in context. In *EMNLP*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018b. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Empirical Methods in Natural Language Processing*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640.

Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. Domain-robust dialogue rewriting as sequence tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Antonios Minas Krasakis, Andrew Yates, and E. Kanoulas. 2022. Zero-shot query contextualization for conversational search. *ArXiv*, abs/2204.10613.

Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2425–2428.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy J. Lin. 2021. Contextualized query embeddings for conversational search. In *EMNLP*.

Hang Liu, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021a. Conversational query rewriting with self-supervised learning. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7628–7632.

Hui Liu, Zhan Shi, and Xiao-Dan Zhu. 2021b. Unsupervised conversation disentanglement through co-training. In *EMNLP*.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. *ArXiv*, abs/2009.13166.

Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, R. Perego, Nicola Tonellotto, and Ophir Frieder. 2020. Topic propagation in conversational search. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019a. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2114–2124.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019b. In *Empirical Methods in Natural Language Processing*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of ACL: EMNLP 2020)*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *NAACL*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Shuangyong Song, Chao Wang, Qianqian Xie, Xinxing Zu, Huan Chen, and Haiqing Chen. 2020. A two-stage conversational query rewriting model with multi-task learning. In *Companion Proceedings of the Web Conference 2020*, pages 6–7.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. *ACL*.

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and A. Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

Svitlana Vakulenko, S. Longpre, Zhucheng Tu, and R. Anantha. 2021a. Question rewriting for conversational question answering. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and S. Longpre. 2021b. A comparison of question rewriting methods for conversational passage retrieval. *ArXiv*, abs/2101.07382.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 921–930.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jiawei Wu, Lei Li, and William Yang Wang. 2018. Reinforced co-training. In *NAACL*.

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. 2020. Semantic role labeling guided multi-turn dialogue rewriter. In *EMNLP*.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1933–1936.

Shih Yuan Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yifei Yuan, Chen Shi, Runze Wang, Liyi Chen, Feijun Jiang, Yuan You, and Wai Lam. 2022. Mcqueen: a benchmark for multimodal conversational query rewrite. *ArXiv*, abs/2210.12775.

Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu. 2020. Filling conversation ellipsis for better social dialog understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9587–9595.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

Zhi-Hua Zhou and Ming Li. 2010. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439.

A. Co-training Algorithm

Algorithm 1 shows the detailed algorithm of our co-training paradigm.

B. Limitations and Future Works

Although our model has shown effectiveness in the CQR task, one drawback is that, the quality of the unlabeled data is vital to the model performance. How to choose hyperparameters such as the confidence threshold of the selectors is important but tricky. In addition, for the page limit, our work focus on the query rewrite mostly on the NLP level, while how much this paradigm will benifit the conversational information retrieval system is still underexplored. Moreover, this Rewriter/Simplifier system can be adapted to other generative tasks, where in this work we only focus on the query write

Algorithm 1 Simplifier and Rewriter Co-training Paradigm

```
Require:
    Simplifier: $S$, Rewriter: $R$
    Labeled dataset: $D$, Unlabeled Rewriter and Simplifier Dataset: $U_R$, $U_S$
    Simplifier Confidence Threshold: $s_s$, Rewriter Confidence Threshold: $s_r$
Ensure:
    A trained Simplifier $S^*$, A trained Rewriter $R^*$
Initialize $S$ and $R$ and train them on $D$
while $U_S \neq \varnothing$ and $U_R \neq \varnothing$ do
    $P_S \leftarrow [\,]$ , $P_R \leftarrow [\,]$
    for $q_s \in U_S$ do
        $q'_r \leftarrow$ Generate the simplified query By $S$
        Compute confidence score $s_x$
        if $s_x > s_s$ then
            $P_S$.insert $(q'_r, q_s, s_x)$
        end if
    end for
    for $q_r \in U_R$ do
        $q'_s \leftarrow$ Generate the rewritten query By $R$
        Compute confidence score $s_y$
        if $s_y > s_r$ then
            $P_R$.insert $(q_r, q'_s, s_y)$
        end if
    end for
    $U_S \leftarrow U_S \backslash P_S$, $U_R \leftarrow U_R \backslash P_R$, $P \leftarrow P_S \cup P_R$
    $D_{Aug} \leftarrow Aug(D, P)$
    Train $S$ and $R$ on $D_{Aug}$
end while
```

task. In our future work, we will work on exploring the co-training paradigm under other scenarios. We'll be exploring how co-training can be applied specifically for other conversational IR scenarios, ultimately enhancing user experiences and satisfaction.